Lessons Learned and Future Directions for the AMBR Model Comparison Project

Kevin A. Gluck
Research Psychologist
Air Force Research Laboratory
6030 S. Kent St.
Mesa, AZ 85212
480-988-6561 x-234
kevin.gluck@williams.af.mil

Richard W. Pew
Principal Scientist
BBN Technologies
Cambridge, MA 02138
617-873-3557
pew@bbn.com

Keywords:
cognitive modeling, human behavior representation, multi-tasking, HLA, concept learning

ABSTRACT: *The first iteration of AFRL's Agent-based Modeling and Behavior Representation (AMBR) Model Comparison Project was quite a learning experience for all involved. This paper focuses on feedback received, challenges faced, and lessons learned during and after Round 1 of the AMBR Model Comparison. We include a section on implications for future (or other) human model comparisons, in the hope that others who may adopt this general methodology will find useful suggestions for planning their own human model comparisons. The paper ends with a description of current plans for AMBR Rounds 3 and 4.*

1. Introduction

This paper is the last of several presented as part of the AMBR Model Comparison Symposium at the 10th Annual Conference on Computer-Generated Forces and Behavior Representation. We mentioned in the introductory paper for this symposium [1] that the first goal of the AMBR Model Comparison Project is to advance the state of the art in cognitive and behavioral modeling. The other Symposium papers provide ample evidence that the participating modeling architectures were challenged and improved as a direct result of their participation in this project, which we consider to be an indication of success in advancing the state of the art.

This positive outcome notwithstanding, we have found that the comparison of human behavior representation (HBR) models is a challenging undertaking. One reason for this is that it is an unusual occurrence. It is rare to have the opportunity to compare and contrast a variety of models created by developers who use different model architectures and draw their models from different theoretical and practical perspectives. There are no clear methodological guidelines for engaging in such comparisons. Despite the challenges involved, it *still* is the case that the AMBR Model Comparison provided a context in which the participating architectures were motivated to expand and improve their capabilities. This suggests that the general design of a comparison of different HBR models to a common set of human performance data is a fruitful one. If it continues to prove fruitful (in later rounds of the project), then hopefully others will be motivated to try this general methodology. If that is to be the case, then we feel a professional responsibility to share our "lessons learned" regarding the planning and implementation of an HBR model comparison.

These lessons are drawn from three sources. First is the expert panel, who met with the AMBR Model Comparison organizers and participants near the end of Round 1. Next is feedback from the modeling teams who have participated in the project so far. Last are our own reflections on these first two rounds of the model comparison, as well as conversations we've had with other interested persons outside the project.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2001	**Conference Proceedings**	**01-01-2000 to 31-04-2001**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Lessons Learned and Future Directions for the AMBR Model Comparison Project	**F33615-99-C-6002**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER **63231F**

6. AUTHOR(S)	5d. PROJECT NUMBER
Kevin Gluck; Richard Pew	**4923**
	5e. TASK NUMBER **04**
	5f. WORK UNIT NUMBER **49230401**

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Air Force Research Laboratory/RHA,Warfighter Readiness Research Division,6030 South Kent Street,Mesa,AZ,85212-6061	**AFRL; AFRL/RHA**

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
Air Force Research Laboratory/RHA, Warfighter Readiness Research Division, 6030 South Kent Street, Mesa, AZ, 85212-6061	**AFRL; AFRL/RHA**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) **AFRL-RH-AZ-PR-2001-0003**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
In Proceedings of the 10th Computer-Generated Forces and Behavior Representation (CGF-BR) Conference, held 14-17 May 01, in Norfolk VA

14. ABSTRACT
The first iteration of AFRL?s Agent-based Modeling and Behavior Representation (AMBR) Model Comparison Project was quite a learning experience for all involved. This paper focuses on feedback received, challenges faced, and lessons learned during and after Round 1 of the AMBR Model Comparison. We include a section on implications for future (or other) human model comparisons, in the hope that others who may adopt this general methodology will find useful suggestions for planning their own human model comparisons. The paper ends with a description of current plans for AMBR Rounds 3 and 4.

15. SUBJECT TERMS
Agent-based modeling and behavior representation; AMBR; Model comparison; Lessons learned; Human model comparisons; Human behavior representation; Behavior modeling; Modeling architectures; Cognitive modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Public Release**	**17**	

2. Expert Panel

Near the completion of Round 1, a panel of experts in human performance model development and evaluation was convened to provide an appraisal of the results of the first round. The panel included the following distinguished individuals:

- **Dr. Sheldon Baron**, BBN Technologies, retired
- **Dr. Wayne Gray**, Associate Professor of Psychology, Human Factors & Applied Cognitive Program, George Mason University
- **Dr. Harold Hawkins**, Program Manager, Office of Naval Research
- **Dr. Peter Polson**, Professor of Psychology, University of Colorado

The panelists volunteered their time and provided valuable feedback and suggestions, for which we are grateful.

The panel was charged with providing an evaluation covering the following topics: (a) critique of AMBR Round 1 design and execution, (b) summary of the strengths and weaknesses of each model, and (c) discuss issues, challenges, and recommendations for future rounds of the project.

2.1 Critique of AMBR Round 1 Design and Execution

2.1.1 Small sample, high variability

Prior to the expert panel meeting, we had collected two sets of human performance data for use in the comparison. One set (N = 8) was provided to the modeling teams for "tuning" their models. The other set (N = 8), collected from different participants doing an analogous set of scenarios, was reserved for the comparison of the models to human performance. This design was adopted partly to prevent "over-tuning" the models to a particular set of comparison data, and partly to provide a test of the robustness of the models. The problem we ran into, and that made it that much more challenging for the panel to assess the "goodness" of the models' predictions, was that there was a fair amount of variability in performance within each data set, and one group of participants was better at the task than others. The small sample and high variability in the "comparison" data set had the panel concerned about the validity of concluding that this sample was an accurate representation of the central tendency of human performance in this task. This of course made it impossible to engage in any sort of head-to-head comparison regarding which of the models had the best "fit" to the data.

After the expert panel meeting, the team revised the data reporting to include both the model development data and the comparison data, combined into a single set. The models were re-run, this time on the development data scenarios, and those runs were aggregated with the other data for comparison with the aggregated human performance data. This modification resulted in a more generalizable representation of the central tendency and variability of human performance in the ATC task. It is these data that are reported in Tenney and Spector [2].

2.1.2 Inadequate assessment of robustness

One rationale for the original 2-dataset design (a "tuning" set and a "comparison" set) was that the different scenarios would serve as an assessment of the robustness of the models. During model development, the modelers did not even have access to the scenarios used for comparison data collection, so they weren't sure what to expect and had to design their models in such a way that they would generalize to a new set of scenarios.

It was made clear, however, that the interface symbology and behavioral requirements would remain consistent across scenario sets. In fact, the only thing that changed was the location and timing of the appearance of planes on the radar screen. The panel noted that this made the comparison scenarios so similar to the tuning scenarios, that they hardly constituted a robustness check at all. There was a silver lining, however, in that the similarities between the two sets of scenarios made it that much more justifiable to combine them for increased sample size.

2.1.3 Limitations of aggregate outcome data

The expert panel noted that all of the human performance data to which we were comparing the models' predictions were aggregate outcome data. No analysis was completed at the level of individual participants, and none of the analysis focused on the micro-processes involved in completing a scenario. The panel found that having human performance and model data only at the level of aggregate outcome measures made it very difficult to distinguish the models on the basis of the fidelity of their predictions.

There were two exceptions to this in Round 1. One is that CHI Systems' CGF-COGNET model made predictions about each of the six workload measures in the TLX, rather than just the aggregate measure. Another is that stochastic performance characteristics of Carnegie

Mellon's ACT-R model made it possible for them to predict the range of variability in the human data, in addition to the central tendency. Those two models distinguished themselves by voluntarily going beyond the minimum prediction requirements of the comparison. Those minimum requirements all involved predictions of the central tendency of aggregate outcome measures

The decision was explicitly made to focus the analyses and comparisons at the aggregate outcome level earlier in the project. This was done knowing full well that others have warned of the dangers of limiting analysis in this manner, extolled the virtues of individual participant analyses, and found idiographic analyses to be informative with respect to cognitive process [3, 4, 5]. Nevertheless, in this case, pragmatic considerations (time, funding) prevailed. The fact is, taking the data analysis to finer levels of detail is more costly and time-consuming and this reality can not be ignored. We simply did not allot sufficient time and resources in Round 1 for idiographic data analysis, and therefore were restricted to the more traditional nomothetic approach.

2.1.4 Incomplete understanding of model implementation

Another challenge the panel encountered in executing its charge was that they were not able to come to a complete enough understanding of each of the models to really feel comfortable in doing an assessment of strengths and weaknesses.

For such an assessment to be accurate, and for it to have any constructive influence, requires a great deal of knowledge of both the underlying modeling architecture and of the specific implementation of that model. In planning the agenda for the review, we allowed approximately two hours of presentation, discussion, and demo time per model. This turned out to be only enough time to gain surface-level familiarity with the models – not enough to accomplish what we initially were asking of the expert panelists.

It was, however, enough time for the talented panel to identify some of the accomplishments of each of these models. These are described in the next section, accompanied by elaborations of our own. Section 2.2 assumes some level of familiarity with the models developed in Round 1, which can be achieved by reading the model papers from this symposium [6, 7, 8, 9].

2.2 Summary of Modeler Accomplishments

2.2.1 ACT-R

The ACT-R team was a late addition to the project. They did not know about AMBR during the initial bidding and awarding of contracts. It wasn't until Harold Hawkins, of the Office of Naval Research, stepped up with funding for the ACT-R team (at about half the level of funding of the other two funded teams) that they committed to participating. This happened approximately four months before the models were due, which put the ACT-R team at a disadvantage regarding model development time. The panel acknowledged that their accomplishments are particularly impressive, given the reduced time and resources.

The incorporation of an explicit representation of sensitivity to visual onsets, which is new for an ACT-R model, allowed for the possibility of task interruptions, and therefore increased reactivity in the model. This is an important milestone for the ACT-R group, because much of the cognitive modeling community had assumed that ACT-R's goal-focused orientation precluded the possibility of task interruptions, thereby limiting the utility of ACT-R as an architecture for modeling multi-tasking. That the addition of sensitivity to visual onsets made this possible serves as additional evidence for the modeling benefits to be gained by using an "embodied" cognitive architecture.

Finally, the panel observed that the ACT-R model's ability to multi-task, as well as its ability to approximate the variability in human performance (described earlier as a characteristic that distinguished this model from the others), were both based on previously-existing symbolic and subsymbolic characteristics of the architecture. This is a good thing, from the standpoint of reusability and generalizability of architectural features. Given that one of the goals of the model comparison is to encourage various architectures to extend themselves in new ways, it was not clear to the panel whether this kind of re-use should necessarily be considered "better" than having developed a new, special-purpose multi-tasking capability for this model. Perhaps the right attitude is simply to consider it noteworthy.

2.2.2 COGNET/iGEN

Whereas the other architectures are motivated by the goal of a unified theory of human cognition and action, the COGNET/iGEN team are quite explicit regarding the fact that the purpose for development of their modeling architecture is not to put forth new theory. Rather, their goal is to develop an expert system shell with practical applicability in a wide variety of modeling contexts. Just as is the case for the theory-motivated architectures, the panel noted that the recent addition of sensory-motor

representations in CGF-COGNET is an important extension of their architecture. It is consistent with the general trend toward "embodiment" in cognitive modeling [10, 11, 12].

The COGNET/iGEN model stood out from the crowd by way of its ability to link each of the six NASA-TLX workload sub-measures to quantitative components of the model. The panel acknowledged this as an accomplishment for the model, and also suggested that the very fact that it was possible serves to increase the construct validity of the TLX sub-measures. So this was a win on two fronts.

Finally, the panel observed that the CGF-COGNET variant used for this model includes the addition of a separate knowledge type (metacognitive knowledge), which in conjunction with declarative and procedural knowledge, enables the model to multi-task. This received mixed reviews from the panel. On the one hand, it clearly makes for an effective means of managing activity during multi-tasking, while on the other hand the panel questioned the theoretical parsimony of a separate metacognitive knowledge mechanism.

2.2.3 D-COG

The panel found AFRL's D-COG model to be an interesting new approach to building a cognitive architecture. The D-COG team were commended for introducing a novel architecture that had the potential to address some new classes of issues. A consequence of creating an architecture with a design that really breaks new ground, is that it is even more difficult than usual to understand that design, because many of the underlying representational and processing assumptions are novel as well. The panel mostly had questions about the D-COG model, and precious few conclusions. It was never clear to the panel exactly how multi-tasking was implemented in D-COG, or how the architecture can be used to arrive at response time predictions.

Although the core constructs were in place, the actual implementation of the D-COG architecture was under development even as the Round 1 model was being developed. This makes it all that much more impressive that the D-COG team managed to get a model completed, and should be considered an accomplishment in itself.

2.2.4 EPIC-Soar

In addition, the EPIC-Soar model developed for AMBR Round 1 broke from the Soar tradition of a single procedural long-term memory store and added a representation of long-term declarative memory that included ACT-R's declarative knowledge decay functions. This synthesis of elements from three different existing architectures was considered by the panel to be an exciting accomplishment. It also is an example of the kind of productive cross-fertilization that can result from bringing different modeling architectures to bear on common human performance challenges.

The EPIC-Soar model includes a "visualization tool" layer that provides real-time information about the current focus of the model's visual attention and cognitive activity as the model is running. The panel was enthusiastic about this as a development, debugging, and demonstration tool.

The EPIC-Soar approach to modeling multitasking consisted of adding a capability for task interruption, but no explicit, separate representation for multitasking per se.

2.3 An Issue, a Challenge, and a Recommendation for Future Rounds

The panel finished off its conclusions with an issue/challenge/recommendation triumvirate that addresses the essence of what we found difficult in Round 1 of the AMBR Model Comparison.

The primary issue we faced throughout the project was *identifying* specific points of comparison among the models. Attempts at identifying specific points of comparison occurred throughout Round 1. These discussions happened at a couple of group meetings scheduled shortly after modeling contracts were awarded, and also by email and phone, and eventually some consensus was reached. Prior to the expert panel review, the Moderator provided a set of topics to the modelers that they were to address in their presentations. These included the following:

- Overview of model, describing unique features
- Theory/architecture on which model is based
- Description of how the model works
- Psychological findings, assumptions and intuitions underlying your model
- Unique challenges of this task and how they were handled
- Approach to model development
- Demo of the model on a common scenario

Going into the panel review, this seemed like a fine list of topics for the modeling team presentations, and each was also intended as a point of comparison among the models. To their credit, the modeling teams all did an admirable job of addressing each point. What we failed to appreciate beforehand was how difficult it was going to be to bring everyone up to speed on the first three topics (unique features, underlying theory, and how the model works) in the allotted time. We mentioned this point earlier, and don't want to belabor it, but it is extremely challenging to get a group of people to a deep level of understanding of four different modeling architectures and the details of four different models built within those architectures – all in two days.

So even with these points of comparison identified, the remaining challenge is to actually *follow through* with the comparison. This requires some real understanding of those models and their underlying architectures, which we did not successfully achieve in such a short time. This is where the process broke down. The entire panel review was spent trying to achieve a sufficient level of understanding, and we never actually managed the direct comparison across models that had been intended.

A recommendation from the panel for addressing this challenge in future rounds of the project is to get panel members involved earlier. If they have more knowledge of the modeling focus, task, participating modeling architectures, and the developed models prior to the panel review, then perhaps a sufficient level of understanding can be achieved more quickly. This should facilitate following through more completely with the direct comparison.

3. Feedback from Modeling Teams

As Round 1 drew to a close, and we began to plan for the subsequent rounds, a solicitation went out to the modeling teams for feedback on Round 1. Their replies identified three areas of concern regarding how Round 1 was designed and implemented.

3.1 Access to Simulation Code and Human Data

A significant point of contention for the modeling teams was the fact that the task code and the human performance data were not available at the time contracts were awarded. Not having access to the task code was frustrating for the modelers, who were interested in beginning to hook their architectures into the simulation code as early as possible. One of the requirements was that the models actually interface with the same task the humans were using during data collection, so it is of course perfectly reasonable that the modelers wanted access to the code. Not having the code frozen and ready for delivery at the time of contract awards was a programmatic oversight that we plan to avoid in the future.

Related to this was the complaint that human performance data were provided rather late in the contract period. Continuing software development, decision making about performance measures, and pilot testing all delayed completion of the simulation code, which of course also delayed delivery of the human performance data. Clearly, the modelers managed to overcome these issues in the end, but the preference for earlier delivery of both simulation code and human data came through clearly in the feedback.

3.2 Lack of a Common CTA

There was some concern (although not unanimous) regarding the fact that there was no centralized cognitive task analysis (CTA) on which model behaviors could be based. This meant that, one way or another, each modeling team had to complete their own CTA. The effect of this is that it creates another source of variance among the models that makes it that much harder to compare them on a case-by-case basis. It adds a knowledge level confound into any direct comparisons among the models, such that in addition to architectural features and theoretical assumptions, it becomes necessary to compare differences in task knowledge and strategies that are built into the models.

3.3 Grain Size of the Performance Analysis

Finally, two points came up in the model team feedback regarding the grain size of the performance analysis. The first was that it would have been useful to have eye movement data from even a small sample of participants. This can help constrain the models and ameliorate concerns associated with the lack of a common CTA.

The second point was the same one that came up in the discussion of feedback from the expert panel, regarding individual vs. aggregate data analysis. Since this concern was discussed in section 2.1.3, we won't revisit the issue in any detail here. The following email excerpt from one of the modelers sums the point up nicely:

"… judging the quality of such a model merely by comparison to summary outcome measures ignores various levels of verisimilitude and associated model utility which could otherwise be examined."

4. Implications for Future Comparisons

To this point, we have shared a variety of challenges from and pieces of feedback regarding the first round of the comparison. What do we conclude from these points, and what are the implications of those conclusions for future comparisons (in the AMBR project or other HBR model comparisons)?

4.1 Programmatic Concerns

One major grouping of the points made earlier in the paper can fall under the heading of "programmatic concerns." These are issues related to the organization, scheduling, and management of the comparison.

4.1.1 Timely access to simulation code and human data

Although we had an intellectual understanding of the importance of timely access to simulation code and human data prior to Round 1, the project agenda in Round 1 did not reflect this understanding. As a result, there was frustration and floundering among the modelers for a while, as the simulation code was being completed, then as the human performance data were collected.

The implication for future comparisons is that delivery of frozen simulation code should occur immediately after the awarding of modeler contracts and delivery of human performance data should follow shortly thereafter. The entire program agenda should be designed with this in mind.

4.1.2 Common CTA

Although one can always make a valid argument for providing a common CTA for a model comparison, we tend to think that the importance of this increases with the complexity of the task and the amount of training required to achieve adequate competence. Thus, to the extent that the simulation environment is more toward the "lab task" end of the continuum, a common CTA is less important.

The jury is still out on this, and deliberations are likely to continue, but we tend to think that future AMBR comparisons also will not involve a common CTA.

4.1.3 Involvement of the expert panelists

We think it is likely that the model comparison would benefit from increased involvement of the panelists, and we intend to get them involved earlier, and on a more continuous basis, throughout future model comparisons. This is likely to take a variety of forms, including (a) participation on the model team selection committee, (b) consultations regarding empirical design and points of model comparison early in each round, (c) a 2-stage review process involving the expert panel in which an initial assessment of the models is made, then the modelers have an opportunity to make improvements to the models before convening for the final comparison, and (d) an extra day added to the end of the final comparison, dedicated to following through on the points of comparison identified earlier in the round.

4.2 Empirical Concerns

A second major grouping of the points made earlier in the paper can fall under the heading of "empirical concerns." These are issues related to the experimental design, data analysis, and model assessment.

4.2.1 Data sources and analyses

We are increasingly convinced that a combination of individual and aggregate data, collected from as wide a variety of sources (computer log files, eye movements, verbal protocols) as is possible given the time and funding constraints of the project, provides the best hope for truly stressing the predictive boundaries of human modeling architectures. Stressing these boundaries is likely to illuminate shortcomings and increase the probability of further advancements in the state of the art. Future AMBR model comparisons are likely to include a wider variety of data, analyzed at both individual participant and aggregate levels.

4.2.2 Model robustness

The scenarios represented in the evaluation data were too similar to those for the development data to allow tests of the robustness of the models. However, improving on this weakness is not straightforward. On the one hand each model was designed to represent specific task requirements. To present scenarios that stretch the model demands beyond the specified task requirements seems "unfair" and, to some extent, uninformative.

It does seem "fair" to specify to the developers the "envelope" within which task parameters might vary, and to use as large an envelope as possible, given the project's time and funding constraints, but not to specify where in that envelope the specific scenarios will be selected for evaluation. Then the development data and the evaluation data could be drawn from different regions of that envelope. It also might be appropriate and insightful to

formulate one "challenge scenario" that takes the modeling requirements judiciously outside the "envelope," but to go beyond that does not seem productive. Exactly how one identifies the boundaries of this envelope, and what it means to be "judiciously outside" the envelope remain unspecified.

4.2.3 Increasing understanding of the models

As discussed earlier, actually following through on some form of head-to-head model comparisons requires first that those making the assessment have a thorough understanding of the implementation of the models. We have already mentioned, in the section on programmatic concerns, that we will strive for increased expert panel participation in future rounds. What else can we do to bring about a deeper understanding of the models for purposes of the comparison? Here are several ideas currently under discussion for future rounds:

(a) A code walk-through was suggested (by the Round 1 panel, in fact) as possibly a useful component of future reviews. Clearly, the utility of this depends on the technical background of the panel members, so the decision whether to do this is primarily up to them.

(b) Characterize each model in terms of the number of fixed and variable parameters. Alternatively, partition the parameters into those that characterize the task and the working environment and those that reflect the human representation, and then describe the subset of the human parameters that were "adapted" to the specifics of the task and development data rather than fixed *a priori*. However, models differ in the extent to which they independently characterize the task environment and the human performance, so that such a partitioning would not always be possible or even sensible.

(c) Require the model developers to undertake sensitivity analyses to relate the fit of their models to the performance data as a function of the setting of one or more critical parameters of the model.

(d) Encourage each modeler to include an interface that makes the internal processing of the model transparent to an observer. The EPIC-Soar team provided a dynamic pictorial representation of the eye scan patterns being undertaken by the model together with status lights that indicated the class of the activity that was being undertaken at each moment in time. These features were very helpful in understanding the sequencing of interruptions and activities that were being undertaken.

(e) Ask the modelers to collaborate with the moderator team to create a comprehensive table of features across models. This structure should be developed early in the project so that it can be used by all the developers to catalog their models at the time of the comparison and evaluation.

We hope this list and the earlier description of Round 1 challenges and feedback will serve as fodder for discussion among others interested in hosting an HBR model comparison. The paper now turns to a brief description of the future of the AMBR Model Comparison.

5. Current Plans for Future Comparisons

Pew and Mavor [13] note that, despite the fact that there is an enormous literature on memory and learning in the experimental psychology, cognitive psychology, and cognitive science fields, and despite the fact that this research is relevant to the representation of human behavior in military simulations, " . . . current military simulations make little or no use of learning models" (p. 148). It is the need for advancements in the state of the art for modeling learning processes that has persuaded us to focus this model comparison on models of *concept learning* in the context of the ATC task.

5.1 Round 3: Concept Learning

The task for Round 3 will retain the multi-tasking perceptual-motor features of the air traffic control (ATC) task used in AMBR Rounds 1 and 2. This task is described in detail in Tenney and Spector [2]. It is still undecided whether the task will retain the HLA-based federation architecture developed for AMBR Round 2 (the Icarus Federation), or the task will return to the non-HLA format.

The significant change in the task from Round 1 to Round 2 will be that in place of the speed query, it will contain an embedded concept learning task. Multiple aircraft will query the controller (the controller that is being modeled) about the possibility of changing altitude. The controller will make a decision to authorize an altitude change based on a multi-dimensional attribute matrix that might include dimensions like aircraft size, level of atmospheric turbulence, and current altitude. The controller must learn the appropriate responses on the basis of feedback received through the user interface concerning whether they made a correct decision or not. This feature structure of this concept learning task is based on the laboratory study by Shepard, Hovland and Jenkins [14], and modeling studies reported by Nosofsky et al. [15].

5.2 Round 4: More on Concept Learning

The task for Round 4 will be fundamentally similar to the task used in Round 3, but the details are still under consideration. Based on the results of the Round 3 model evaluations, the Round 4 task will be designed to further stress the models and examine their capabilities (modeling team contracts will extend through both Rounds 3 and 4). We anticipate a focus on the ability of models to adapt from one set of learned concepts to a new, changed set of concepts based on the same or a similar set of concept attributes. Other manipulations such as the workload of the perceptual motor task may also be explored as deemed appropriate given the results of Round 3.

6. Acknowledgements

The content of this paper was influenced by the summary of the expert panel report presented by Wayne Gray at the 44th Annual Meeting of the Human Factors and Ergonomics Society, and by BBN's final report for AMBR Round 1, to which Yvette Tenney, Steve Deutsch, Sandy Spector, and Brett Benyo contributed. The opinions expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Air Force, the U.S. Department of Defense, or the U.S. Government.

7. References

[1] Gluck, K. A., & Pew, R. W. (2001). Overview of the Agent-based Modeling and Behavior Representation (AMBR) Model Comparison Project. *Proceedings of the 10th Annual Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA.

[2] Tenney, Y. J., & Spector, S. L. (2001). Comparisons of HBR Models with Human-in-the-loop Performance in a Simplified Air Traffic Control Simulation with and without HLA Protocols: Task Simulation, Human Data and Results. *Proceedings of the 10th Annual Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA.

[3] Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*(3), 250-264.

[4] Gobet, F., & Ritter, F. E. (2000). Individual data analysis and unified theories of cognition: A methodological proposal. *Proceedings of the 3rd International Conference on Cognitive Modeling*. Veenendaal: Universal Press.

[5] Gluck, K. A., Staszewski, J. J., Richman, H., Simon, H. A., & Delahanty, P. (submitted). The right tool for the job: Information-processing analysis in categorization. *23rd Annual Meeting of the Cognitive Science Society*, Edinburgh, Scotland.

[6] Lebiere, C., Anderson, J. R., & Bothell, D. (2001). Multi-tasking and cognitive workload in an ACT-R model of a simplified air traffic control task. *Proceedings of the 10th Annual Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA.

[7] Zachary, W., Santarelli, T., Ryder, J., Stokes, J. & Scolaro, D. (2001). Developing a multi-tasking cognitive agent using the COGNET/iGEN integrative architecture. *Proceedings of the 10th Annual Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA.

[8] Eggleston, R. G., Young, M. J., & McCreight, K. L. (2001). Modeling human work through distributed cognition. *Proceedings of the 10th Annual Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA.

[9] Chong, R. (2001). Low-level behavioral modeling and the HLA: An EPIC-Soar model of an enroute air-traffic control task. *Proceedings of the 10th Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA.

[10] Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12*, 391-438.

[11] Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson & C. Lebiere, *The atomic components of thought* (pp. 167-200). Mahwah, NJ: Erlbaum.

[12] Chong, R. S., & Laird, J. E. (1997). Identifying dual-task executive process knowledge using EPIC-Soar. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 107-112). Mahwah, NJ: Erlbaum.

[13] Pew, R. W., & Mavor, A. S. (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington, D. C.: National Academy Press.

[14] Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13, Whole No. 517).

[15] Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: a replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*, 352-369.

Author Biographies

KEVIN GLUCK is a research psychologist at the Air Force Research Laboratory's Warfighter Training Research Division in Mesa, AZ. He is the program manager for AFRL's AMBR Model Comparison Project, and AFRL's POC for development and testing of the Icarus Federation. Dr. Gluck earned a B.A. in Psychology from Trinity University in 1993, an M.S. in Cognitive Psychology from Carnegie Mellon University in 1997, and a Ph.D. in Cognitive Psychology from Carnegie Mellon University in 1999.

RICHARD PEW is Principal Scientist at BBN Technologies LLC, a unit of the Verizon Technology Organization in Cambridge, Massachusetts. Dr. Pew holds a bachelors degree in Electrical Engineering from Cornell University (1956), a master of arts degree in Psychology from Harvard University (1960) and a PhD in Psychology with a specialization in Engineering Psychology from The University of Michigan (1963). Throughout his career he has been involved in the development and utilization of human performance models and in the conduct of experimental and field studies of human performance in applied settings. He spent 11 years on the faculty of the Psychology Department at Michigan before moving to BBN in 1974. His current research interests include the impact of automation on human performance, human-computer interaction and human performance modeling. He is the author or co-author of over 90 book chapters, research papers, conference proceedings, and technical reports.

Lessons Learned and Future Directions for the AMBR Model Comparison

15 May 2001

Kevin A. Gluck
Air Force Research Laboratory

Richard W. Pew
BBN Technologies

Overview

- HBR Accomplishments
- Feedback from Expert Panel
- Feedback from Modelers
- Implications for Future Comparisons
- AMBR Rounds 3 and 4

HBR Accomplishments

- **ACT-R**
  - Unit task framework for multi-tasking
  - Visual onsets/color changes
  - Workload computation
- **COGNET/iGEN**
  - First application of CGF (i.e., embodied) architecture
  - Detailed workload computation
- **D-COG**
  - Opportunity to build and test the architecture
- **EPIC-Soar**
  - Most dynamic/complex application to date
  - Declarative base-level activation and decay
  - Workload computation
- **All**
  - Transferred model to HLA federation

Feedback from Expert Panel

- **Small sample, large individual differences**
- **Limited utility of aggregate outcome data**
- **Weak assessment of robustness**
- **Incomplete understanding of model implementations**

Feedback from Modelers

- **Want earlier access to simulation code and human data**
- **Possibly beneficial to have a common CTA**
- **Grain-size of the behavior analysis**

Implications for Future Comparisons

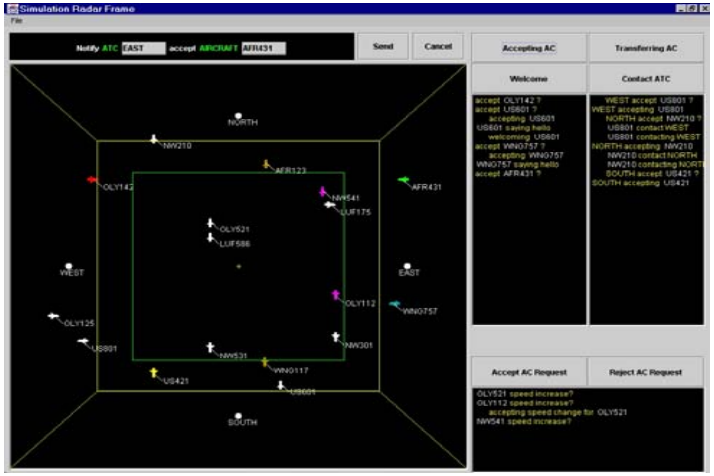
- **Programmatic Changes**
  - Earlier access to simulation and data (*definitely*)
  - Increased involvement of the expert panelists (*definitely*)
  - Generate common CTA (*under consideration*)

- **Empirical Changes**
  - Aggregate and individual analyses (*definitely*)
  - Increase understanding of the models (*definitely*)
  - High-density data (*probably*)
  - Stronger test of model robustness (*under consideration*)

AMBR Rounds 3 and 4

Modeling Focus

Category Learning



Context

Modified version of ATC task

Controller must learn to respond appropriately to requests for altitude change

Moderator

BBN (Pew, Deutsch, Diller, Tenney, Spector, Benyo)

Modelers

{To be selected}

Questions?

Suggestions?

Comments?

(all are welcome)